Learning Video Representations from Large Language Models

Year: 2022 | *Citations: 225* | *Authors: Yue Zhao, Ishan Misra, Philipp Krahenbuhl*

Abstract

We introduce LAVILA, a new approach to learning video-language representations by leveraging Large Language Models (LLMs). We repurpose pre-trained LLMs to be conditioned on visual input, and finetune them to create automatic video narrators. Our auto-generated narrations offer a number of advantages, including dense coverage of long videos, better temporal synchronization of the visual information and text, and much higher diversity of text. The video-language embedding learned contrastively with these narrations outperforms the previous state-of-the-art on multiple first-person and third-person video tasks, both in zero-shot and finetuned setups. Most notably, Lavilaobtains an absolute gain of 10.1% on EGTEA classification and 5.9% Epic-Kitchens-100 multi-instance retrieval benchmarks. Furthermore, LaVilatrained with only half the narrations from the Ego4D dataset outperforms models trained on the full set, and shows positive scaling behavior on increasing pre-training data and model size.